ISSN (Print) 1225-9918
ISSN (Online) 2287-3406
Journal of Life Science 2018 Vol. 28. No. 11. 1255~1261 DOI : https://doi.org/10.5352/JLS.2018.28.11.1255

Imputation Accuracy from 770K SNP Chips to Next Generation Sequencing Data in a Hanwoo (Korean Native Cattle) Population using Minimac3 and Beagle

Na-Rae An, Ju-Hwan Son, Jong-Eun Park, Han-Ha Chai, Gul-Won Jang and Dajeong Lim*

Animal Genomics and Bioinformatics Division, National Institute of Animal Science, Rural Development Administration, Wanju 565-851, Korea

Received June 4, 2018 /Revised November 9, 2018 /Accepted November 16, 2018

Whole genome analysis have been made possible with the development of DNA sequencing technologies and discovery of many single nucleotide polymorphisms (SNPs). Large number of SNP can be analyzed with SNP chips, since SNPs of human as well as livestock genomes are available. Among the various missing nucleotide imputation programs, Minimac3 software is suggested to be highly accurate, with a simplified workflow and relatively fast. In the present study, we used Minimac3 program to perform genomic missing value substitution 1,226 animals 770K SNP chip and imputing missing SNPs with next generation sequencing data from 311 animals. The accuracy on each chromosome was about 94~96%, and individual sample accuracy was about 92~98%. After imputation of the genotypes, SNPs with R Square (R^2) values for three conditions were 0.4, 0.6, and 0.8 and the percentage of SNPs were 91%, 84%, and 70% respectively. The differences in the Minor Allele Frequency gave R^2 values corresponding to seven intervals (0, 0.025), (0.025, 0.05), (0.05, 0.1), (0.1, 0.2), (0.2, 0.3). (0.3, 0.4) and (0.4, 0.5) of 64~88%. The total analysis time was about 12 hr. In future SNP chip studies, as the size and complexity of the genomic datasets increase, we expect that genomic imputation using Minimac3 can improve the reliability of chip data for Hanwoo discrimination.

Key words : Hanwoo, imputation, Minimac3, Next Generation Sequencing (NGS), SNP chip

서 론

과학 및 분석 기술의 발전으로 20세기 후반부터 유전과 변이에 관여된 염색체와 유전자 발현을 중심으로 연구하는 분자유전학의 발전을 가져왔다. 따라서 이러한 유전학을 이용하는 축산분야에도 많은 변화를 가져오게 되었다. 더불어 분자유전학은 기존의 통계 중심의 선발육종방식을 DNA염기서열을 활용한 마커 도움 선발(maker assisted selection)로 변화 가능케 하였다. 또한 DNA 염기서열분석의 발전과 많은 단일염기다형성(Single Nucleotide Polymorphism; SNP)의 발굴은 마커 도움 선발을 더욱 더 정확하게 할 수 있도록 만들었다[20, 21]. 하지만 SNP의 유전자형을 분석하기 위해서는 분석 시간과 비용이 많이 소요된다. 이러한 유전자형분석에서의 단점을 보완하고자 유용 SNP마커를 모아 하나의 칩으로 제작하였다. 따라서 SNP칩의 개발은 유전체선발에 있어서 유전자형분석에 대한 시간적 부분에서의 획기적인 변화를 가져오게 되었다. 계속되는 염기서열분석과 발굴되는 SNP마커들로 인해 더욱 더 용량이 큰 칩의 제작을 필요로 하여 고밀도 SNP칩이 개발되었다. 하지만 고밀도 SNP칩은 비용이 매우 비싸서 마커 도움 선발을 하는데 있어 경제적인 문제를 발생시켰다. 이런 비용적 문제점을 해결하기 위해 유전자형 대치법(imputation)이 개발되었다[10, 16, 25]. 유전자형 대치법은 유전자형 마커 데이터를 참조하여 전장유전체연관분석(Genome‐Wide Association Study; GWAS)의 데이터에 적용해서 누락되어 채워지지 않은 유전자형의 정보를 추정하는 것을 말한다[12]. 이는 서로 다른 칩 데이터를 하나의 데이터로 통합하는 경우에서도 유용하게 쓰이며, 고밀도의 칩 데이터 정보를 whole genome 데이터로 확대 및 대치하는 경우에도 쓰이고 있다. 저밀도의 칩을 사용하면서도 고밀도의 칩을 사용한 효과를 나타내도록 하는 장점이 있으며, 다른 품종의 유전자형을 추정하거나 비용을 감소시키는 효과가 있다[4]. 유전자형 대치법에 대한 분석을 위하여 fastPHASE [26], Findhap [30], AlphaImpute [11], 그리고 Fimpute [24] 등이 있다. 유전자형의 imputation 정확도에 영향을 미치는 요소에는 참조 집단의 구조 및 크기, 유전자형 빈도와 연관 불평형(Linkage Disequilibrium; LD), 품종 간의 차이[5], 분석 프로그램 등 여러 요소의 영향을 받는다[5, 13, 16].

기존 연구에서는 유전자형과 표현형을 활용한 결측치 대치(imputation)를 통해 육종가추정(EBV)이 가능하며[29], 양에서는 Illumina 50K ovine 칩의 54,977개의 SNP 데이터와 프로그램 fastPHASE와 Beagle을 이용한 결측치 대치 정확도를 측정한 결과 86~96%의 정확도를 얻었다[1, 10, 26]. 앵거스

*Corresponding author
Tel : +82-63-238-7306, Fax : +82-63-238-7347
E-mail : lim.dj@korea.kr

(Angus) 소에서도 Illumina BovineSNP50 Bead 칩 데이터를 이용하여 SNP 유전자형에 대해 결측치 대치 추정을 한 결과 87~99%의 정확도를 산출하였다. 염색체 1번의 정확도의 경우 Beagle (v3.3)을 이용하였을 때 0.986으로 가장 높았고, Alpha Impute를 이용하였을 때 0.908로 가장 낮았다. 또한 염색체 16번의 경우 Beagle을 이용하였을 때 0.984로 가장 높았으며, AlphaImpute를 이용하였을 때 0.909로 가장 낮았다. 염색체 28번의 경우 역시 Beagle을 이용하였을 때 0.971로 정확도가 가장 높았고, Findhap (v2)를 이용하였을 때 0.868로 가장 낮았다[27]. Beagle의 경우 모든 동물이 관련이 없다는 가정하에서도 parent-offspring trio와 parent-offspring pairs 패키지를 처리할 수 있기 때문에 세 개의 염색체 모두 가장 높은 정확도가 나타났다. 1,366두 일본 흑우의 연구에서는 저밀도 SNP칩인 Illumina BovineSNP50 Bead 칩의 38,502개의 SNP 데이터와 Beagle (v3.3.2) 프로그램을 이용한 결측치 대치 정확도에 대한 결과는 최대 97%였다[23]. 또한 고밀도 SNP칩인 Illumina BovineHD Bead 칩을 사용하여 50K 칩으로 1,368두 일본 흑우의 Beagle (v4) [2]을 사용한 연구에서의 결측치 대치 정확도를 추정한 결과 약 99%의 정확도를 보였다[28]. 결측치 대치의 정확도에 대한 연구는 식물의 경우 3K 또는 35K의 SNP 칩이 주로 사용되며[12, 14], 양은 50K 칩[10], 소는 50K, 770K 칩이 주로 사용된다[8, 29]. 그러나 한우(Hanwoo)는 칩 데이터를 이용하여 SNP 유전자형을 대치한 결과에 대한 정확도 분석은 아직 이루어진 바 없다. 따라서 본 연구는 Human 1,000 genomes project에서 대규모 genotyping project 자료를 처리하기 위해 만들어진 소프트웨어인 Eagle (v2.3.2)과 Minimac3 (v2.0.1)를 사용하여 한우 770K 칩 데이터를 이용하여 차세대 염기서열분석 데이터로 결측치 대치의 정확도를 조사하였으며, Beagle (v4.1)과 Minimac3의 결측치 대치 정확도의 비교를 통해 Minimac3의 높은 정확도에 대해 확인해 보았다.

재료 및 방법

SNP Chip 분석

본 연구에서 사용 된 데이터 세트는 SNP 칩으로 유전자형이 결정된 한우 735,293개의 SNP 마커가 포함된 1,226두의 Illumina Bovine 770K (HD 칩)와 27,081,079개의 SNP 마커가 포함된 311두의 차세대 염기서열분석 데이터를 사용하였다 (Table 1).

유전자형 및 전장유전체 분석

1,226두의 770K 칩 VCFtools (v.0.1.13) [6]의 vcf 파일과 311두의 차세대 염기서열분석 데이터의 vcf 파일을 이용하였다. 이 데이터들 중에 서로 공통인 개체 50두를 770K 칩에서는 대상(target) 파일로 사용하고 차세대 염기서열분석 데이터에서는 이 공통된 개체 50두를 제외한 나머지 261두의 차세대 염기서열분석 데이터를 참조(reference) 파일로 이용하였다.

Table 1. Number of Single Nucleotide Polymorphism (SNP) in 770K chip-seq and Next Generation Sequencing (NGS)

Chromosome	SNP	
	770K	NGS
chr1	46,495	1,729,789
chr2	40,056	1,361,664
chr3	35,579	1,319,183
chr4	34,980	1,360,455
chr5	34,842	1,236,597
chr6	35,519	1,281,025
chr7	33,168	1,161,369
chr8	33,529	1,151,008
chr9	31,060	1,066,767
chr10	30,449	1,105,732
chr11	32,015	1,071,565
chr12	26,127	1,108,017
chr13	23,594	897,284
chr14	24,780	865,334
chr15	24,755	982,399
chr16	24,178	880,315
chr17	22,266	794,264
chr18	19,386	699,171
chr19	18,908	682,186
chr20	21,490	762,598
chr21	21,175	765,797
chr22	18,034	633,500
chr23	15,215	750,842
chr24	18,620	711,162
chr25	12,931	473,884
chr26	15,242	567,042
chr27	13,152	508,155
chr28	13,038	517,275
chr29	14,710	636,700
Total	735,293	27,081,079

품질관리(Quality Control)

대상 파일과 참조 파일에 대해 VCFtools의 명령어를 사용하여 대립유전자형빈도(Minor Allele Frequency; MAF)는 0.05 이하, 하디-와인버그 평형(Hardy-Weinberg Equilibrium; HWE)은 10^{-6} 이하인 SNP를 제외하는 품질관리(Quality Control; QC)를 진행하였다.

결측치 대치(Imputation)

QC가 완료된 후, 파일들을 1번부터 29번까지의 염색체 별로 분리하였다. 염색체 별로 나누어진 대상 파일과 참조 파일의 vcf 파일을 두가지 방법으로 결측치 대치를 하였다. 한가지 방법은 Eagle (v.2.3.2) [18]을 이용하여 위상(phase)을 분석하여, 위상 분석이 완료된 후 Minimac3 (v.2.0.1) [7]를 이용하여 유전자형 결측치 대치를 하였다. 또 다른 방법은 Beagle (v4.1)

[3]를 이용하여 유전자형 결측치 대치를 하였다.

정확도 분석

차세대 염기서열분석 데이터의 공통 개체 50두 vcf 파일과 Minimac3와 Beagle 프로그램을 각각 사용하여 결측치 대치가 완료된 파일들을 비교하여, 정확도를 분석하였다. Minimac3와 Beagle의 유전자형 결측치 대치의 정확도를 염색체별, 개체별로 비교하였고, Minimac3를 사용한 결측치 대치 결과 정확도에 따른 SNP의 개수와 염색체별 R Square (R^2) 값, 염색체별 결측치 대치 완료 시간 및 대립유전자형빈도 값에 대한 R^2 값에 대해서도 분석하였다.

결 과

정확도별 SNP Chip 분석 결과

770K 칩 데이터에서 차세대 염기서열분석 데이터로 Minimac3와 Beagle을 각각 사용한 유전자형의 대치(Genotype imputation) 결과 중 정확도별 SNP의 개수는 Table 2에 나타내었다. Minimac3를 사용하여 대치가 완료된 총 14,021,408개의 SNP의 염색체별 정확도는 0.955로 나타났으며, 모든 개체에서 최소 정확도가 0인 SNP는 1,627개였다. 반면에 모든 개체에서 최고 정확도 1로 나타난 SNP의 개수는 5,356,134개였다. 그 밖에 SNP 개수가 1~10% 미만인 SNP는 281개였으며, 10~20% 사이의 SNP는 3,560개, 20~30% 사이의 SNP는 11,418개, 30~40% 사이의 SNP는 23,137개, 40~50% 사이의 SNP는 48,142개였다. 따라서 10~50%에 해당하는 정확도는 10%의 사이마다 약 2배 이상의 SNP의 개수가 증가 되었음을 보였다. 50~ 60%사이의 SNP의 수는 65,859개이며, 60~70% 사이의 SNP의 수는 122,445개, 70~80% 사이의 SNP의 수는 282,967개, 80~ 90% 사이의 SNP의 수는 807,294개였다. 그리고 가장 많은 SNP의 개수 7,298,544개를 차지한 정확도는 90~99%였다.

Table 2. Accuracy (in %) of the imputed SNPs with Minimac3 & Beagle

Accuracy (%)	Number of SNPs	
	Minimac3	Beagle
0	1,627	1,534
1~10	281	2,226
10~20	3,560	7,326
20~30	11,418	41,622
30~40	23,137	186,354
40~50	48,142	464,532
50~60	65,859	823,312
60~70	122,445	1,227,182
70~80	282,967	1,767,509
80~90	807,294	2,478,735
90~99	7,298,544	2,301,959
100	5,356,134	1,991,351
Total	14,021,408	11,293,642

Beagle을 이용하여 대치가 완료된 총 11,293,642개의 SNP의 염색체별 정확도는 0.805로 나타났으며, 모든 개체에서 최소 정확도가 0인 SNP는 1,534개였다. 반면에 모든 개체에서 최고 정확도 1로 나타난 SNP의 개수는 1,991,351개였다. 그 밖에 SNP 개수가 1~10% 미만인 SNP는 2,226개였으며, 10~20% 사이의 SNP는 7,326개, 20~30% 사이의 SNP는 41,622개, 30~40%사이의 SNP는 186,354개, 40~50% 사이의 SNP는 464,532개, 50~60% 사이의 SNP의 수는 823,312개였다. 60~70% 사이의 SNP의 수는 1,227,182개, 70~80% 사이의 SNP의수는 1,767,509개, 80~90% 사이의 SNP의 수는 2,301,959개, 90~99% 사이의 SNP의 수는 2,301,959개였으며, 이중에서 가장 많은 SNP의 개수를 차지한 정확도는 80~90%였다.

염색체별 유전자형의 정확도 분석 결과

유전자형의 결측치가 대치된 1번 염색체부터 29번 염색체까지의 염색체별 정확도는 Table 3과 Fig. 1에 나타냈다. 염색체별 정확도의 평균은 Beagle과 Minimac3를 이용한 경우에는 각각 0.805와 0.955로 나타났다. Beagle로 분석한 결과 23번 염색체가 0.784로 가장 낮은 정확도를 보였으며, 1번 염색체가 0.817로 가장 높은 정확도를 가졌고, Beagle의 염색체별 정확도는 78~82%의 정확도 분포를 보였다. 반면에 Minimac3로 분석한 결과 18번 염색체가 0.940로 가장 낮은 정확도를 보였으며, 2번 염색체가 0.964로 가장 높은 정확도를 가졌고, 염색체별 정확도는 94~96%의 정확도의 분포를 보였다.

개체별 정확도 분석 결과

Minimac3를 이용하여 결측치 대치가 완료된 파일과 차세대 염기서열분석 데이터 파일의 동일 개체 50두의 개체별 정확도를 Fig. 2에 나타내었다. 50두의 개체의 평균 정확도는 0.955로 나타났다. 가장 낮은 정확도는 0.923이며, 가장 높은

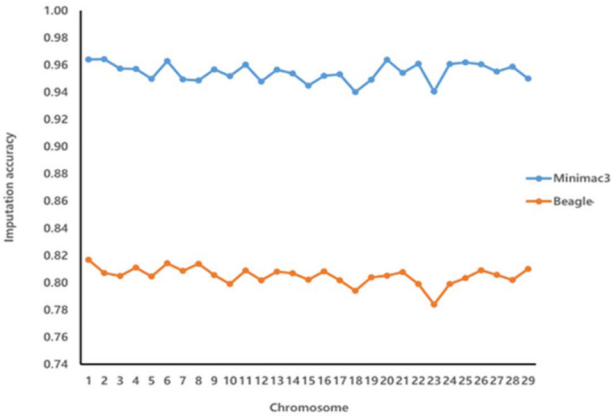

Fig. 1. Line plot showing the accuracy of imputation for each chromosome using Minimac3 & Beagle. X axis is chromosome number, Y axis is imputation accuracy.

Table 3. Chromosome wise average accuracy (in %) using Minimac3 & Beagle

Chromosome	Average accuracy (%)	
	Minimac3	Beagle
chr 1	0.96(96.41%)	82(81.69%)
chr 2	0.96(96.44%)	81(80.73%)
chr 3	0.96(95.73%)	80(80.50%)
chr 4	0.96(95.71%)	81(81.12%)
chr 5	0.95(94.99%)	80(80.47%)
chr 6	0.96(96.29%)	81(81.43%)
chr 7	0.95(94.94%)	81(80.88%)
chr 8	0.95(94.87%)	81(81.39%)
chr 9	0.96(95.67%)	81(80.56%)
chr 10	0.95(95.18%)	80(79.92%)
chr 11	0.96(96.02%)	81(80.90%)
chr 12	0.95(94.79%)	80(80.19%)
chr 13	0.96(95.66%)	81(80.82%)
chr 14	0.95(95.38%)	81(80.71%)
chr 15	0.94(94.49%)	80(80.22%)
chr 16	0.95(95.21%)	81(80.83%)
chr 17	0.95(95.32%)	80(80.18%)
chr 18	0.94(94.01%)	79(79.41%)
chr 19	0.95(94.93%)	80(80.40%)
chr 20	0.96(96.00%)	81(80.53%)
chr 21	0.95(95.42%)	81(80.78%)
chr 22	0.96(96.10%)	80(79.90%)
chr 23	0.94(94.05%)	78(78.40%)
chr 24	0.96(96.07%)	80(79.90%)
chr 25	0.96(96.19%)	80(80.34%)
chr 26	0.96(96.05%)	81(80.93%)
chr 27	0.96(95.51%)	81(80.57%)
chr 28	0.96(95.88%)	80(80.20%)
chr 29	0.95(95.00%)	81(81.02%)
Total	0.96(95.47%)	80(80.51%)

정확도는 0.970이였다. 동일 개체의 분포를 살펴보면 92~93% 정확도에 해당하는 개체는 3두였으며, 93~94% 정확도에 해당되는 개체 또한 3두였다. 94~95% 정확도에 해당하는 개체는 9두이며, 95~96% 정확도에 해당되는 개체는 11두였다. 가장 많은 22두가 해당되는 정확도는 96~97%였으며, 가장 적은 2두가 해당되는 정확도는 97~98%로 나타났다.

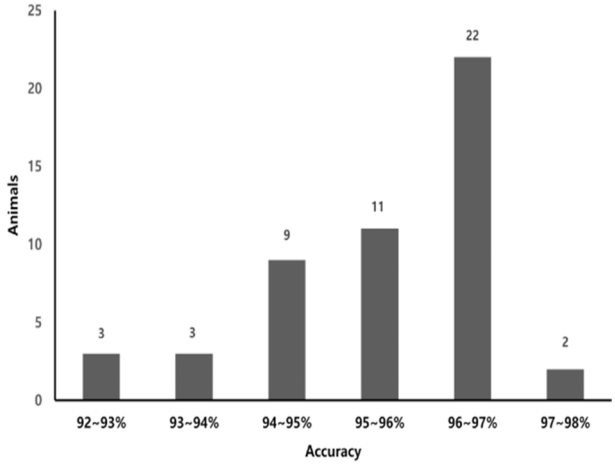

Fig. 2. Bar plot showing the accuracy of imputation of 50 Hanwoo bulls that were genotyped from both whole genome sequence data and 777K SNP chip data. X axis is percentage of accuracy, Y axis is number of animals.

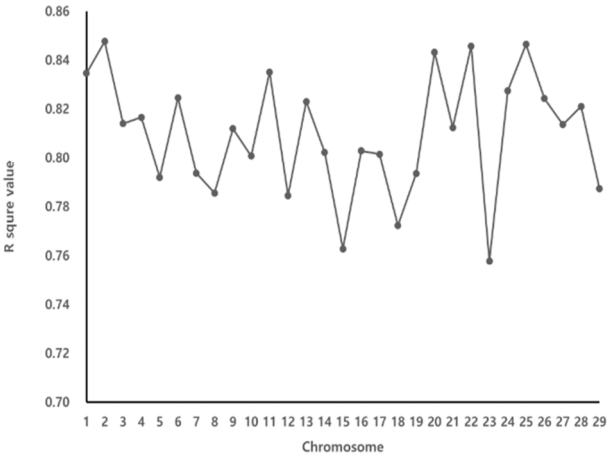

Fig. 3. Effect of chromosome size on imputation accuracy calculated through analyzing the correlation (R^2) between chromosome size & accuracy of imputation. X axis is chromosome number, Y axis is R^2.

염색체별 R^2 값 분석 결과

염색체별 R^2 값에 대한 그래프를 Fig. 3에 나타냈다. Minimac3의 R^2 값은 하디-와인버그 평형에서 예상되는 이항 변이에 대한 대립 유전자 양의 관측된 분산 비율이다. R^2 값의 분포는 0.75~0.85 값 사이에 존재하였다. 23번 염색체가 0.758로 가장 낮은 값을 보였으며, 2번 염색체가 0.848로 가장 높은 값을 보였다. Minimac3 프로그램을 사용하여 유전자형의 결측치 대치가 완료된 파일의 총 SNP의 개수는 12,314,008개였다. 이 중 R^2 값이 0.4 이상인 SNP의 개수는 11,149,352개로 총 SNP의 91%였고, R^2 값이 0.6 이상인 SNP는 10,394,106개로 84%였으며, R^2 값이 0.8 이상인 SNP는 8,642,910개로 70%였다.

대립유전자형빈도(Minor Allele Frequency) 값에 대한 R^2 값 분석 결과

마지막으로 MAF 값에 대한 R^2 값의 그래프를 Fig. 4에 나타냈다. 유전자형의 대체가 완료된 50두의 SNP에 대해 MAF의 차이를 기준으로 (0, 0.025), (0.025, 0.05), (0.05, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), (0.4, 0.5)의 7구간에 해당하는 R^2 값을 계산하였다. MAF가 0~0.2 사이일 때 R^2 값이 큰 폭으로 증가했다.

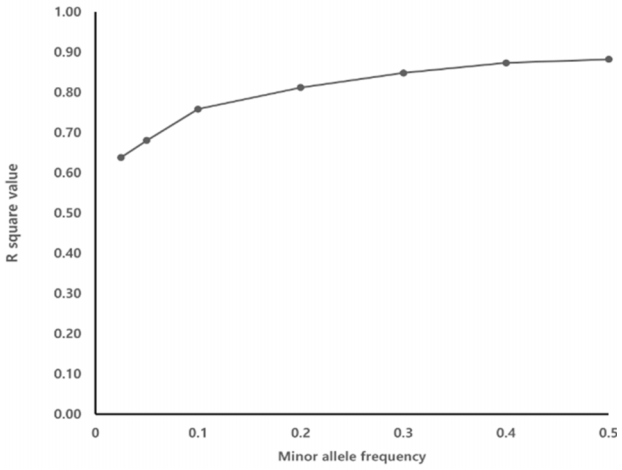

Fig. 4. Effect of MAF (Minor Allele Frequency) on imputation accuracy calculated through analyzing the correlation (R^2) between MAF & accuracy of imputation. X axis is MAF, Y axis is R^2.

고 찰

유전자형 대치법을 통해서 유전자형을 추정하는 것은 경제적 측면에서 매우 필요하다. 본 연구에서는 한우의 저밀도 칩을 바탕으로 고밀도 칩의 차세대 염기서열분석 데이터를 추정하여 그 정확도를 산출하는 작업을 수행하였다. 결측치 대치 분석 자주 쓰이는 프로그램으로 Beagle, fastPHASE는 연관불평형 정보를 사용하여 누락 된 SNP 유전자형을 설명하고[16, 20], Sun 등[27]에 따르면 Findhap, AlphaImpute, Fimpute는 혈통 정보와 연관불평형 정보를 모두 이용하여 유전자형의 결측치를 대체 할 수 있었다. 또한 이들의 앵거스 연구에서 Beagle, Impute, fastPHASE, findhap, AlphaImpute 그리고 Fimpute 프로그램을 각각 사용하여 유전자형 결측치를 추정하여 정확도를 얻었다. 그 결과 Beagle 프로그램을 사용하여 결측치 대치를 하는 것이 정확도가 가장 높게 나왔다. 또한 Beagle은 multi-allelic markers의 처리가 가능하며, 따라서 Beagle은 lager reference panels의 가장 적합한 프로그램으로 유전자형 결측치 대치법으로 많이 사용되었다[8]. Beagle은 염색체 위치 정보를 이용하여 참조 패널을 유사한 영역별로 분리한 후, 일배체형(haplotype)의 유사성을 기반으로 국소적 일배체형 클러스터링(localized haplotypes clustering)을 통해 Hidden Markov Models (HMM)을 수행한다[1]. 하지만 지역에 따라 다양하게 클러스터가 존재하므로 메모리 적인 문제와 참조 집단과 샘플 크기에 의해서 계산 시간이 오래 걸리는 단점을 가지고 있다[15, 20, 27].

이에 반해 Minimac3 프로그램은 한우의 SNP 칩 데이터를 활용한 유전자형 대치법의 정확도 연구에서 사용한 프로그램 Minimac3는 서열화된 일배체형 사이의 국소적인 유사성을 이용한다는 점에서는 Beagle과 유사하나, 정확성의 손실 없이 분석을 크게 단순화하고 계산의 효율성을 증가시켜 유전자형의 신속한 대치를 실행한다. Minimac3 프로그램은 누락 된 데이터가 있어도 구조적으로 효율적이며, 더 복잡한 분석에도 정확하고 동일한 결과를 산출해 낸다는 알고리즘의 두 가지 중요한 특징을 가진다. 그리고 Beagle보다 2배 빠른 처리 속도를 가지며, Beagle의 최대 단점인 메모리의 사용량 또한 Minimac3가 Beagle보다 72% 감소시킨다는 이점이 있다[6]. 또한 Fig. 3에서 볼 수 있듯이 R^2 값이 평균 0.81로, 분석 데이터의 적합도(goodness of fit)도 완벽한 적합도에 1에 가까운 높은 적합도를 나타냈다. 따라서 본 연구에서는 Minimac3 프로그램을 이용하여 유전자형 대치법의 정확도를 추정하였다. 유전자형 대치법의 연구 결과 Minimac3를 사용한 결측치 대치의 정확도는 Table 3에서 볼 수 있듯이 약 96%의 정확도를 나타냈다. 이 결과를 토대로 유전자형 결측치 대치를 위한 소프트웨어 Minimac3는 분석 단계를 단순화하여 컴퓨팅 처리 속도를 빠르게 하고(Table 4) 적은 메모리를 사용한다는 측면

Table 4. Chromosome wise time taken by Minimac3 for imputation

Chromosome	Time
chr1	48 min 20 sec
chr2	39 min 41 sec
chr3	37 min 50 sec
chr4	40 min 37 sec
chr5	36 min 41 sec
chr6	37 min 45 sec
chr7	31 min 49 sec
chr8	33 min 17 sec
chr9	30 min 44 sec
chr10	31 min 58 sec
chr11	31 min 28 sec
chr12	29 min 25 sec
chr13	24 min 27 sec
chr14	25 min 21 sec
chr15	29 min 10 sec
chr16	24 min 25 sec
chr17	24 min 10 sec
chr18	20 min 42 sec
chr19	19 min 24 sec
chr20	21 min 54 sec
chr21	22 min 11 sec
chr22	17 min 46 sec
chr23	20 min 32 sec
chr24	19 min 9 sec
chr25	14 min 40 sec
chr26	16 min 7 sec
chr27	15 min 1 sec
chr28	15 min 55 sec
chr29	18 min 4 sec
Total	12 hr 5 min 3 sec

에서 Beagle보다 유용한 프로그램이라고 생각한다.

더 나아가 Minimac3 프로그램을 사용하여 참조 집단의 크기를 달리하거나 다양한 축종의 SNP 칩을 이용한 결측치 대치 정확도를 추정하는 추가적인 연구 등도 진행되어야 한다. 유전체 데이터 세트의 크기와 복잡성이 증가하는 추후의 유전체 SNP 칩 연구에서 Minimac3를 사용한 유전체 결측치 대치법은 한우의 판별에 있어서 칩 데이터의 신뢰도를 향상 시킬 수 있는 프로그램으로 점차 일반화 될 것으로 전망한다. 위의 결과를 토대로 볼 때 Minimac3를 이용한 유전자형 대치법은 한우에서 유전자형을 추정하는 데 있어서 Table 3과 Table 4에서 볼 수 있듯이 높은 정확도를 비롯해 빠른 처리 속도(약 12시간)를 토대로 매우 유용한 프로그램으로 사용될 가능성을 보였다.

감사의 글

본 연구는 농촌진흥청 연구사업(PJ01251902) 연구비를 지원 받아 수행하였습니다.

References

1. Browning, B. L. and Browning, S. R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210-223.
2. Browning, B. L. and Browning, S. R. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471.
3. Browning, B. L. and Browning, S. R. 2016. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116-126.
4. Browning, S. R. and Browning, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084-1097.
5. Chud, T. C., Ventura, R. V., Schenkel, F. S., Carvalheiro, R., Buzanskas, M. E., Rosa, J. O., de Alvarenga Mudadu, M., da Silva, M. V. G., Mokry, F. B. and Marcondes, C. R. 2015. Strategies for genotype imputation in composite beef cattle. *BMC Genet.* **16**, 99.
6. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T. and Sherry, S. T. 2011. The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158.
7. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R. and Fuchsberger, C. 2016. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284-1287.
8. Druet, T., Schrooten, C. and De Roos, A. P. W. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J. Dairy Sci.* **93**, 5443-5454.
9. Ellinghaus, D., Schreiber, S., Franke, A. and Nothnagel, M. 2009. Current software for genotype imputation. *Hum. Genomics* **3**, 371.
10. Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W. and Van der Werf, J. H. J. 2012. Accuracy of genotype imputation in sheep breeds. *Anim. Genet.* **43**, 72-80.
11. Hickey, J. M., Kinghorn, B. P., Tier, B., Wilson, J. F., Dunstan, N. and Van der Werf, J. H. J. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol*. **43**, 12.
12. Hickey, J. M., Crossa, J., Babu, R. and de los Campos, G. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci.* **52**, 654-663.
13. Howie, B. N., Donnelly, P. and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529.
14. Iwata, H. and Jannink, J. L. 2010. Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. *Crop Sci.* **50**, 1269-1278.
15. Johnston, J., Kistemaker, G. and Sullivan, P. G. 2011. Comparison of different imputation methods. Interbull Bulletin. pp. 25-33. Stavanger, Norway.
16. Li, L., Li, Y., Browning, S. R., Browning, B. L., Slater, A. J., Kong, X., Aponte, J. L., Mooser, V. E., Chissoe, S. L., Whittaker, J. C., Nelson, M. R. and Ehm, M. G. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* **6**, e24945.
17. Li, Y., Willer, C., Sanna, S. and Abecasis, G. 2009. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387-406.
18. Loh, P. -R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R. and L Price, A. 2016. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443-1448.
19. Lopes, F. B., Wu, X. L., Li, H., Xu, J., Perkins, T., Genho, J., Ferretti, R., Tait Jr, R. G., Bauck, S. and Rosa, G. J. M. 2018. Improving accuracy of genomic prediction in Brangus cattle by adding animals with imputed low density SNP genotypes. *J. Anim. Breed. Genet.* **135**, 14-27.
20. Marchini, J. and Howie, B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499-511.
21. Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-1829.
22. Ni, G., Cavero, D., Fangmann, A., Erbe, M. and Simianer, H. 2017. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* **49**, 8.

23. Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., Takasuga, A., Sugimoto, Y. and Iwaisaki, H. 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low density panels in Japanese Black cattle. *Anim. Sci. J.* **87**, 3-12.
24. Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. *J. Dairy Sci.* **94**, 421.
25. Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478.
26. Scheet, P. and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629-644.
27. Sun, C., Wu, X. L., Weigel, K. A., Rosa, G. J. M., Bauck, S., Woodward, B. W., Schnabel, R. D., Taylor, J. F. and Gianola, D. 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. *Genet. Res.* **94**, 133-150.
28. Uemoto, Y., Sasaki, S., Sugimoto, Y. and Watanabe, T. 2015. Accuracy of high density genotype imputation in Japanese Black cattle. *Anim. Genet.* **46**, 388-394.
29. VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., Sonstegard, T. S., Connor, E. E., Winters, M., Van Kaam, J. B., Valentini, A., Van Doormaal, B. J., Faust, M. A. and Doak, G. A. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* **96**, 668-678.
30. VanRaden, P. M., O'Connell, J. R., Wiggans, G. R. and Weigel, K. A. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* **43**, 10.

초록 : Minimac3와 Beagle 프로그램을 이용한 한우 770K chip 데이터에서 차세대 염기서열분석 데이터로의 결측치 대치의 정확도 분석

안나래 · 손주환 · 박종은 · 채한화 · 장길원 · 임다정*
(국립축산과학원 동물유전체과)

DNA 염기서열의 발전과 많은 단일염기서열변이 정보(Single Nucleotide polymorphism, SNP)의 발굴은 유전분석을 가능하게 만들었다. 단일염기서열변이 정보가 사람의 유전체뿐만 아니라 가축의 유전체에서도 이용할 수 있게 됨에 따라서 SNP 칩 마커를 통해 유전자형의 분석이 가능하게 되었다. 여러 유전자형 대치프로그램 중에서도 Minimac3 소프트웨어는 비교적 정확성이 높고, 계산의 효율성을 위해 분석을 단순화하여 유전자형의 결측치 대치 분석 시간을 단축시킨다. 따라서 본 연구에서는 Minimac3 프로그램을 사용하여 한우 1,226두 770K SNP 칩 데이터와 311두 차세대 염기서열분석 데이터를 이용하여 유전자형 결측치 대치를 실행해 보았다. 그 결과 염색체별 정확도는 약 94~96%의 정확도를 나타냈으며, 개체별 정확도는 약 92~98%의 정확도를 나타냈다. 유전자형의 결측치 대치의 완료 후, R Square (R^2) 값이 0.4 이상인 SNP는 총 SNP의 약 91%였다. R^2 값이 0.6 이상인 SNP는 84%였으며, R^2 값이 0.8 이상인 SNP는 70%였다. 대립유전자형빈도 차이를 기준으로 (0, 0.025), (0.025, 0.05), (0.05, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), (0.4, 0.5)의 7구간에 해당하는 R^2 값은 64~88%였다. 결측치 대치의 총 분석 시간은 약 12시간이 걸렸다. 추후의 유전체 데이터 세트의 크기와 복잡성이 증가하는 SNP 칩 연구에서 Minimac3를 사용한 유전체 결측치 대치법은 한우의 판별에 있어서 칩 데이터의 신뢰도를 향상 시킬 수 있을 것으로 본다.